МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Уральский федеральный университет имени
первого Президента России Б. Н. Ельцина»

МАКРОСТАТИСТИЧЕСКИЙ АНАЛИЗ И ПРОГНОЗИРОВАНИЕ ДАННЫХ

Лекция № 5

Прогнозирование данных

Екатеринбург
2024

Содержание

Часть 1. Понятие прогнозирования

Прогнозирование (forecasting) – научный, основанный на установленных причинно-следственных связях и закономерностях исследуемого временного ряда, расчет состояния и вероятностных путей развития явлений и процессов, лежащих в основе ВР.

При прогнозировании мы пытаемся по тем свойствам, характеристикам и состояниям ВР, что были получены в ходе его анализа, описать и дать характеристику процессов и явлений в будущем. Чаще всего прогнозы классифицируют по времени упреждения будущих событий: на **краткосрочные**, **среднесрочные** и **долгосрочные** прогнозы.

Краткосрочные прогнозы прогнозируют исходный ряд на 2-3 отсчета вперед. То есть, если это был ряд среднемесячных чисел, то для него краткосрочным прогнозом будет считаться прогноз на 2-3 месяца, если же ряд содержал годовые значения, то краткосрочный прогноз будет длиной в 2-3 года. То есть «срочность» прогноза всегда определяется его периодом выборки. **Среднесрочные прогнозы** обычно прогнозируют ряд на один сезон или один цикл вперед, то есть в зависимости от внутренней структуры ряда этот интервал может меняться. **Долгосрочные прогнозы** обычно предсказывают значения ряда на несколько циклов/сезонов, либо же на один цикл, имеющий самый крупный временной период.

Общий алгоритм прогнозирования ВР может быть сведен в несколько этапов: анализ исходного временного ряда прогнозирования, выбор метода прогнозирования, построение исходной модели прогноза, ее реализация и проверка достоверности, точности и обоснованности прогноза. В целом схема этого алгоритма приведена на рисунке 5.1.

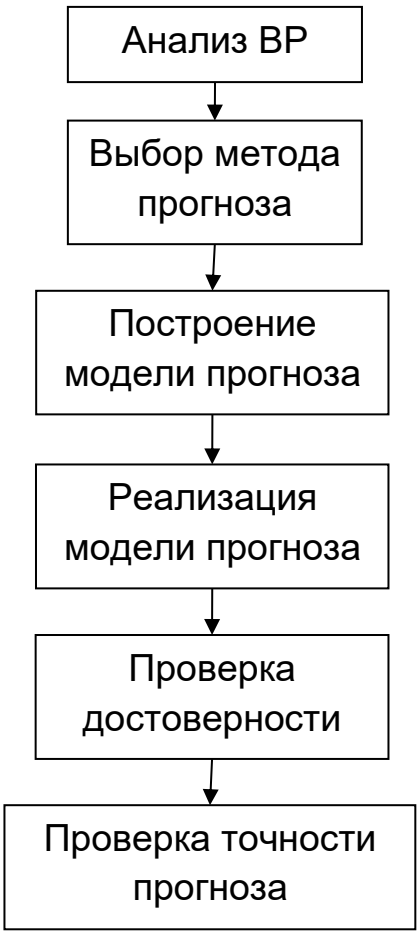

Рисунок 5.1 – Схема общего алгоритма этапов прогнозирования

Самым простым способом прогнозирования множества точек вообще является процесс **экстраполяции**. Под экстраполяцией следует понимать продление тенденций и закономерностей, зафиксированных в прошлом и настоящем, на будущие значения. Здесь выделяют четыре простых метода: прогноз при неизменных уровнях ряда, метод среднего уровня ряда, метод абсолютного прироста и метод среднего темпа роста.

Часть 2. Простейший прогноз на основе экстраполяции

Самый простой прогноз заключается в предположении, что во ВР предшествующие средние уровни вообще не изменяются. Такой прогноз на **M** шагов вперед математически будет выглядеть как:

$$\overline{y}_{t+1} = \overline{y},\ \overline{y}_{t+M} = \overline{y}_{t+M-1} \tag{5.1}$$

то есть мы будем просто продолжать линию мат. ожидания. Понятно, что такой прогноз вряд ли можно считать сложным, но для начального приближения в краткосрочной перспективе иногда даже такой прогноз будет необходим.

При прогнозировании на основе **среднего уровня**, выражение (5.1) опять используется для построения прогноза, но теперь мы еще и оцениваем точность прогноза с помощью понятия **доверительного интервала**. **Интервал** называется **доверительным**, если относительно него можно с заранее выбранной вероятностью утверждать, что он содержит значение прогнозируемого показателя. Доверительный интервал образует две границы: верхнюю h_B и нижнюю h_H.

$$\begin{aligned} h_B &= g - \Delta, \\ h_H &= g + \Delta, \end{aligned} \tag{5.2}$$

где g – оценочный средний прогноз, Δ – доверительный интервал.

Для метода среднего уровня, прогноз с учетом доверительного интервала оценивается достаточно просто:

$$y_{t+M} = \overline{y} \pm t_\alpha \sigma_y \sqrt{1 + \frac{1}{N}} \tag{5.3}$$

где t_α – значение критерия Стьюдента, $\sigma_y = \sqrt{\frac{\sum (y_t - \overline{y})^2}{N-1}}$, N – длина ряда.

2.1. Метод прогноза среднего абсолютного прироста

При прогнозировании методом **среднего абсолютного прироста**, используется следующая методика:

1) Проверяются абсолютные приросты $\nabla y = y_t - y_{t-1}$. Они должны быть примерно одинаковыми.

2) Проверяется неравенство вида $\sigma^2 \le \rho^2$, где $\sigma^2 = \frac{\sum\left(y_t - \bar{y}\right)^2}{N}$ и $\rho^2 = \frac{1}{2}\frac{\sum\left(\nabla y\right)^2}{N}$.

3) Если проверки выполняются, то модель прогноза выглядит, как:

$$y_{t+M} = y_t + M \cdot \frac{y_N - y_1}{N-1}. \qquad (5.4)$$

2.2. Метод прогноза среднего темпа роста

При прогнозировании методом **среднего темпа роста**, используется следующая формула:

$$y_{t+M} = y_t \cdot \bar{T}_p^M, \qquad (5.5)$$

где $\bar{T}_p$ – средний темп роста, вычисляемый по формуле $\bar{T}_p = \left(y_N / y_1\right)^{\frac{1}{N-1}}$.

Этот метод применяется, если темпы роста оказываются не аддитивными, а цепными, то есть остаются примерно одинаковыми, но при этом тенденция развития ряда подчиняется геометрической прогрессии и может быть описана показательной кривой. Проще говоря, если ряд имеет тренд, сообразный экспоненциальной кривой, то для данного ряда такой экстраполяционный прогноз вполне может сработать.

Часть 3. Прогнозирование тренда временного ряда

В предыдущих лекциях обсуждался вопрос выделения тренда из временного ряда различными методами. Эти классы методов делились на две группы: регрессионные методы и сглаживающие методы. К ним еще третьей группой, в принципе, можно добавить изученные модели АРПСС, где тренд описывается в виде авторегрессионной зависимости. Про прогноз на основе моделей авторегрессии мы будет говорить на следующей лекции, здесь же тренд прогнозировать придется либо сглаживанием, либо через регрессию.

При прогнозировании тренда сглаживанием возможны две ситуации. Если сглаживание ряда происходило **экспоненциально**, то прогноз такого тренда происходит по формуле:

$$\tau_{l+1} = (1-\chi)\tau_l + \chi y_l,\ l = N+1, N+2, ..., N+L\ , \qquad (5.6)$$

где χ – коэффициент экспоненциального сглаживания. То есть для прогноза берется последняя точка тренда и отсчета ВР и затем их весовая комбинация даст нам прогноз следующей точки.

Если же сглаживание ряда происходило другим методом, то тогда для получившегося сглаженного тренда необходимо построить сначала «подгоночную» кривую с помощью регрессионных методов.

При прогнозе на основе регрессионных кривых, линейных и нелинейных, после построения модели тренда в виде $y = X\beta + \varepsilon$, прогноз производится на основе выражения:

$$\tau_l = \beta_0 + \beta_1 t_l + \beta_2 t_l^2 + ... + \beta_k t_l^k,\ l = N+1, N+2, ..., N+L\ . \qquad (5.7)$$

Это выражение представлено для линейной регрессии порядка k, но аналогичные выражения прогноза могут быть получены из (5.7) подстановкой вместо степенных зависимостей соответствующих нелинейных функций, которые описывают поведение данной тенденции в рамках выборки временного ряда.

Как ни странно, но прогнозировать тренд оказывается очень легко, в случае его описания на основе регрессионного анализа. Вот только любой прогноз не является абсолютно достоверным, поэтому при прогнозировании тренда недостаточно только описать его будущие отсчеты, нужно еще оценить границы, в которых он будет достоверным.

Для этого используется понятие **доверительного интервала прогноза**. Доверительный интервал всегда определяет две границы – верхнюю и нижнюю. Внутри получившегося «коридора» находится расчетный прогноз. Этот прогноз (будучи случайной величиной) не выходит статистически за границы доверительного интервала с заданной величиной α.

Пусть прогноз тренда строится на основе его модели $\tau(t)$. Границы доверительного интервала описываются выражениями:

$$\begin{aligned} \tau_B(t) &= \tau(t) + \delta(t), \\ \tau_H(t) &= \tau(t) - \delta(t), \end{aligned} \tag{5.8}$$

где $\tau(t)$ – оценка прогноза тренда, $\tau_B(t)$ – верхняя граница доверительного интервала, $\tau_B(t)$ – нижняя граница доверительного интервала.

Тогда нас интересует оценка величины $\delta(t)$. В зависимости от величины порядка полиномиального тренда, расчет $\delta(t)$ будет меняться. Рассмотрим два базовых случая: линейный тренд и квадратичный тренд.

Для линейного тренда, при прогнозе на шаг *l* имеем:

$$\delta_{p=1}(t_l) = t_{\alpha,N-2} \cdot S \cdot \sqrt{1 + \frac{1}{N} + \frac{\left(\tau(t_l) - \overline{\tau}\right)^2}{\sum_{i=1}^{N}\left(\tau_i - \overline{\tau}\right)^2}}, \qquad (5.9)$$

где $S = \sqrt{\frac{\sum_{i=1}^{N} e_i^2}{N-2}}$, e_i – остаточный ряд или ряд ошибок, то есть разница между исходным ВР и его моделью, $t_{\alpha,N-2}$ – квантиль распределения Стьюдента для доверительной вероятности α со степенями свободы $N-2$. Значение квантиля находится по таблицам, либо с помощью специальных функций математических и других расчетных пакетов.

Для квадратичного тренда, при прогнозе на шаг *l* имеем:

$$\delta_{p=2}(t_l) = t_{\alpha,N-2} \cdot S \cdot \sqrt{1 + \frac{\tau(t_l)^2}{\sum_{i=1}^{N}\tau_i^2} + \frac{\sum_{i=1}^{N}\tau_i^4 - 2\cdot\tau(t_l)^2\sum_{i=1}^{N}\tau_i^2 + N\tau(t_l)^4}{N\sum_{i=1}^{N}\tau_i^4 - \left(\sum_{i=1}^{N}\tau_i^2\right)^2}}. \qquad (5.10)$$

Для более высоких степеней эти формулы будут все больше и больше разрастаться, поэтому их мы здесь не приводим. Для степеней полиномов порядка больше 2 либо используют табличные расчетные значения, либо используют специализированные программные функции, либо используют оценки 2 или 1 порядка.

Если построенная модель прогноза адекватна, то с вероятностью α можно утверждать, что при сохранении сложившихся закономерностей развития тенденции, прогноз попадает внутрь доверительного интервала.

Часть 4. Оценка точности и надежности прогнозов

Полученные прогнозы всегда немаловажно оценить по точности и надежности, так как никакой прогноз никогда не является абсолютно достоверным. Эмпирической мерой точности прогноза служит величина его ошибки, которая определяет разность между прогнозным и фактическим значением отсчета ряда. Данный подход работает только в двух случаях: значения прогнозируемых отсчетов уже наблюдаемы, и с ними можно сравнить полученные результаты; либо сразу строится **ретроспективный прогноз**. При **ретроспективном прогнозе** рассчитываются отсчеты ВР для периода времени, за который уже имеются фактические значения. То есть ретроспективный прогноз никогда не выходит за временные рамки, на которых заданы отсчеты анализируемого ВР. Такой прогноз делается с целью проверки схемы прогнозирования.

При ретроспективном прогнозе часть исходного ряда (примерно 2/3) выделяется для анализа и построения прогнозирующих моделей. Оставшаяся часть (около 1/3) тогда используется для оценки ошибок ретроспективного прогноза. Похожим образом работает механизм обучения нейронных сетей – часть выборки идет на обучение, а часть – на проверку.

Любые показатели оценки точности прогноза делятся на три группы:

1) аналитические показатели точности прогноза;
2) сравнительные показатели точности прогноза;
3) качественные показатели точности прогноза.

Рассмотрим каждый из них более подробно.

4.1. Аналитические показатели точности прогноза

Аналитические показатели точности прогноза вычисляют **количественную** величину **ошибки** прогноза. Ошибки прогноза можно вычислять либо по **абсолютному** значению, либо по **относительному**.

Абсолютная ошибка прогноза Δ^* вычисляется, как:

$$\Delta^* = \left| y_t - y_F \right| \tag{5.11}$$

где y_t – фактическое значение отсчета, y_F – прогнозное значение отсчета временного ряда.

Относительная ошибка прогноза $\Delta^*_{отн}$ может быть определена как отношение абсолютной (5.11) ошибки к фактическому значению:

$$\Delta^*_{отн} = \frac{\Delta^*}{y_t} = \frac{\left| y_t - y_F \right|}{y_t} \cdot 100\%, \tag{5.12}$$

либо к прогнозному значению:

$$\Delta^*_{отн} = \frac{\Delta^*}{y_F} = \frac{\left| y_t - y_F \right|}{y_F} \cdot 100\%. \tag{5.13}$$

Абсолютная и относительная ошибки прогноза являются первичной оценкой точности, так как, по сути, являются оценкой точечного прогноза. Точность всей прогнозной модели такими характеристиками будет не оценить.

Поэтому на практике иногда определяются не ошибку прогноза, а некоторый **коэффициент качества прогноза**, который показывает соотношение между числом совпавших отсчетов и общим числом прогнозируемых отсчетов *M*:

$$K_K = \frac{M_{совп}}{M} \tag{5.14}$$

Данный показатель имеет существенный недостаток: что в формуле (5.14) считать *совпавшим отсчетом*? Любая прогнозная точка ряда не будет совпадать с фактическим значением на 100%. Отсюда приходится определять дополнительно некоторую границу близости фактических отсчетов и прогнозных, чтобы считать их совпавшими. Понятно, что чем более широко

трактуется понятие совпадения, тем лучше будет коэффициент качеств, то есть показатель оказывается достаточно субъективным.

Гораздо лучше общим показателем точности прогноза на M точек являются величины **среднего показателя точности прогноза** и **средняя квадратичная ошибка прогноза**. Первый показатель рассчитывается по формуле (здесь $y(t_i)$ - фактические отсчеты ВР, y_i - прогнозные отсчеты):

$$\overline{\Delta}^* = \frac{\sum_{i=1}^{M} \Delta_i^*}{M} = \frac{\sum_{i=1}^{M} \left| y(t_i) - y_i \right|}{M} \tag{5.15}$$

а второй – по формуле:

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^{M} \left(y(t_i) - y_i \right)^2}{M}} \tag{5.16}$$

Обе эти ошибки являются интегральными (то есть обобщающими) показателями точности прогноза, но имеют и общий недостаток – их зависимость от масштаба измерения уровней отсчетов ряда.

Поэтому на практике чаще применяется не абсолютный, а относительный показатель, называемый **средней ошибкой аппроксимации**, который выражается в процентах и рассчитывается как:

$$\overline{\varepsilon} = \frac{1}{M} \sum_{i=1}^{M} \frac{\left| y(t_i) - y_i \right|}{y(t_i)} \cdot 100\% . \tag{5.17}$$

Этот показатель обычно интерпретируется следующим образом: если $\overline{\varepsilon} < 10\%$, то точность прогноза высокая, если $10\% < \overline{\varepsilon} < 20\%$, то точность прогноза хорошая, если $20\% < \overline{\varepsilon} < 50\%$, то точность прогноза удовлетворительная, для $\overline{\varepsilon} > 50\%$ прогноз не удовлетворителен.

4.2. Сравнительные и качественные показатели точности прогноза

В качестве **сравнительного показателя** точности прогноза используется **коэффициент корреляции** между прогнозными и фактическими значениями отсчетов ВР. Получается, что тогда сравнительный показатель рассчитывается по формуле:

$$R = \frac{\frac{1}{M}\sum_{i=1}^{M}\left(y(t_i)-\bar{y}\right)\left(y_i-\bar{y}_F\right)}{\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(y(t_i)-\bar{y}\right)^2 \cdot \frac{1}{M}\sum_{i=1}^{M}\left(y_i-\bar{y}_F\right)^2}}, \qquad (5.18)$$

Используя данный коэффициент в оценке точности прогноза, следует помнить, что коэффициент парной корреляции (5.18), в силу своей сущности, отражает линейное соотношение коррелируемых величин и характеризует лишь взаимосвязь между рядом фактических отсчетов и рядом прогнозных отсчетов. Даже если коэффициент корреляции R=1, то это еще не означает совпадения фактических и прогнозных оценок, а свидетельствует лишь о наличии *линейной зависимости* между прогнозным и фактическим ВР.

Другим показателем оценки точности прогноза может служить **коэффициент несоответствия**, который был предложен Г. Тейлором. Существует три его модификации. Первая модификация:

$$KH_1 = \sqrt{\frac{\sum_{i=1}^{M}\left(y_i - y(t_i)\right)^2}{\sum_{i=1}^{M} y(t_i)^2}}, \qquad (5.19)$$

для которого: если $KH_1 = 0$, то есть полное совпадение фактических $y(t_i)$ и прогнозных y_i отсчетов; если $KH_1 = 1$, то точность соотносима с простейшими методами экстраполяции ряда; если $KH_1 > 1$, то такой прогноз дает результаты хуже, чем, если бы мы просто предполагали неизменность среднего значения ряда.

Вторая модификация рассчитывается как:

$$KH_2 = \sqrt{\frac{\sum_{i=1}^{M}\left(y_i - y(t_i)\right)^2}{\sum_{i=1}^{M}\left(\overline{y} - y(t_i)\right)^2}}, \qquad (5.20)$$

то есть представляет отношение средней квадратичной ошибки прогноза к сумме квадратов отклонений от среднего уровня за весь период прогноза.

Наконец, третья модификация рассчитывается как:

$$KH_3 = \sqrt{\frac{\sum_{i=1}^{M}\left(y_i - y(t_i)\right)^2}{\sum_{i=1}^{M}\left(y(t_i) - \tau_i\right)^2}}. \qquad (5.21)$$

Этот коэффициент отражает отношение средней квадратичной ошибки прогноза к сумме квадратов отклонений фактических отсчетов от отсчетов прогноза по простому тренду. По сути, такой показатель отражает, насколько используемый прогноз лучше или хуже прогноза, который происходил просто на основе экстраполяции тренда исходного ряда.